Published in final edited form as:

Arch Sex Behav. 2010 August ; 39(4): 842–860. doi:10.1007/s10508-009-9505-7.

Consistency of Self-Reported Sexual Behavior in Surveys

Deven T. Hamilton and

Department of Sociology, University of Washington, Seattle, WA, USA

Department of Medical Education and Biomedical Informatics, University of Washington, Box 357240, Seattle, WA 98195-7240, USA

Martina Morris

Department of Sociology, University of Washington, Seattle, WA, USA

Department of Statistics, University of Washington, Seattle, WA, USA

Abstract

Accurate data on sexual behavior have become increasingly important for demographers and epidemiologists, but self-reported data are widely regarded as unreliable. We examined the consistency in the number of sexual partners reported by participants in seven population-based surveys of adults in the U.S. Differences between studies were quite modest and much smaller than those associated with demographic attributes. Surprisingly, the mode of survey administration did not appear to influence disclosure when the questions were similar. We conclude that there is more consistency in sexual partnership reporting than is commonly believed.

Keywords

Sexual behavior; Epidemiology; Partner number; Sex surveys

Introduction

For most of the last century, demographers were the only social scientists that routinely and unobtrusively collected data on sexual behavior in population based surveys. The importance of fertility in the demographic paradigm provided the justification needed to conduct empirical studies of this sensitive topic. With the rapid emergence of incurable life threatening sexually transmitted infections (STIs) like HIV, there is now a broader interest in sexual behavior research to support public health. As a result, data on sexual behavior are now collected more often and in more disciplinary contexts. Once a rare and controversial topic of study, sex has become sufficiently mainstreamed that it is now a routine item in surveys like the General Social Survey (GSS) and the National Health and Nutrition Examination Survey (NHANES).

The renewed scientific interest in this information has again focused attention on the reliability and validity of self-reported sexual behavior. Data on sensitive behaviors are generally regarded as prone to various forms of misreporting (Weinhardt et al., 1998a) and self reports of sexual behavior are thought to have particularly low reliability and validity, especially in the context of risk for HIV (Brody, 1995). Embarrassing or socially stigmatized behaviors like sex are subject to a wide range of reporting problems in surveys (Bancroft, 1997; Bradburn, 1983; Catania, Gibson, Chitwood, & Coates, 1990; Fenton, Johnson, McManus, & Erens, 2001; Weinhardt, Forsyth, Carey, Jaworski, & Durant, 1998b).

The quality of sexual behavior data is difficult to assess because the private nature of the subject is not conducive to direct observation or proxy measures, in contrast to abortion, for example, where provider records can be used to cross-validate self-reports (Fu, Darroch, Henshaw, & Kolb, 1998). While there is no “gold standard” that can be used to evaluate the accuracy of reports, self-administered questionnaires (SAQ) are widely thought to elicit better data (Gribble, Miller, Rogers, & Turner, 1999; Tourangeau & Smith, 1996). If participants are reluctant to admit to another person that they have engaged in illegal or otherwise embarrassing activities, they may be more forthcoming if they can instead disclose it on a confidential form. Studies have also shown that audio computer assisted self-interviewing and computer assisted self-interviewing generally yield higher levels of reporting than does computer assisted personal interviews across a range of items involving sexual behavior and drug use (Jones, 2003; Metzger et al., 2000; Tourangeau & Smith, 1996; Turner et al., 1998).

Another potential influence on reporting is question sequence and response format. Prior items in a survey have been shown to affect responses to later items, since participants tend to eliminate information from their responses if the information has already been captured through previous questions (Schwarz, Strack, & Mai, 1991; Tourangeau, Rasinski, & Bradburn, 1991). For example, if a respondent has been asked about partners in the last year, he or she may subtract those partners in answering subsequent questions about partners in the lifetime. The format of response coding can also affect reporting. When the responses are coded in categories, for example, women tend to report greater numbers of sexual partners if the codes are geared toward higher numbers (Smith, 1992; Tourangeau & Rasinski, 1988). When responses are coded as a continuous variable, participants tend to heap estimates around multiples of 5 or 10, and for values above 20 around 80% of responses are rounded this way (Huttenlocher, Hedges, & Bradburn, 1990; Morris, 1993).

The reliability of self-reports of sexual behavior have been studied for many populations, including gay men (McLaws, Oldenburg, Ross, & Cooper, 1990; Saltzman, Stoddard, McCusker, Moon, & Mayer, 1987), heterosexual men and women (Durant & Carey, 2002; Taylor, Rosen, & Lieblum, 1994; Van Duynhoven, Nagelkerke, & Van de Laar, 1999; Weinhardt et al., 1998a), different racial groups (Sneed et al., 2001), and the mentally ill (Carey, Carey, Maisto, Gordon, & Weinhardt, 2001; Sohler, Colson, Meyer-Bahlburg, & Susser, 2000). The test-retest correlations have been as low as .3 and as high as .9 across these various populations (Catania, Binson, van der Straten, & Stone, 1995). Given the range of correlations, Durant and Carey (2002) suggested that the assessment of sexual behavior might be uniquely difficult to report or particularly difficult to elicit.

While there is a general expectation that participants are reluctant to divulge sensitive information, there is no consensus on the magnitude of misreporting, the variation in estimates across surveys, or the relative impact of the various survey design components on reporting of the number of partners. It is possible that partner number reporting may not be prone to the same biases because it does not elicit the same self-presentation concerns as more taboo sexual behaviors or activities which might be illegal. The purpose of this article is to begin to answer these questions by conducting a systematic comparison of the available empirical data.

We focused on seven population-based surveys that solicited data on the number of sexual partners a person has had in the last year and/or over their lifetime: the General Social Survey (GSS), National College Health Risk Behavior Survey (NCHRBS), National Health and Social Life Survey (NHS LS), National Survey of Family Growth (NSFG), National Survey of Men (NSM), National Survey of Women (NSW), and the Behavioral Risk Factor Survey (BRFS). These surveys were not identical in purpose, design or implementation. However, the samples utilized in these surveys were all population based and they were sufficiently similar that a comparison of their results could shed light on the consistency of sexual behavior reporting and the impact of survey design on the level of disclosure.

There has been some previous work evaluating the sensitive data in a few of these surveys. Jones and Forrest (1992) compared abortion rates reported as part of the NSFG to abortion rate data collected from abortion clinics. The results suggested that fewer than half of all abortions were reported in the NSFG.

This analysis is the first time that these seven surveys have been brought together for comparison. Without a gold standard, we are not in a position to establish the validity of these data. We can, however, show the extent to which responses differ across surveys, after controlling for sample demographic composition, response category coding, year of data collection, and mode of survey administration. This can help to identify the magnitude and sources of variation in self-reported sexual behavior data and provide some guidelines for good study design.

Method

Surveys

Table 1 provides a general summary of the data sets used in this comparative analysis: the GSS, NCHRBS, NHSLs, NSFG, NSM, NSW and the BRFS. There were many differences among these surveys. Some of them had a single, simple question to capture the number of partners, others asked for categories of partners separately (e.g., male and female), while others used a life history calendar approach. There were explicit differences in the sample design and implementation, question sequence, and response categories. These differences required adjustments to permit comparability. In some cases, it was possible to find a lowest common denominator; in other cases, it was not. When the differences proved insurmountable, we conducted separate analyses utilizing subsets of comparable studies. Here, we provide a brief description of these data sets along with a detailed description of the adjustments we made to each (Appendix A provides the specific wording of each

question and a general overview of the primary adjustments). In general, we used a simple rule to guide decisions regarding data adjustment and variable construction: for all surveys, we sought to capture the maximum number of distinct partners reported by participants.

General Social Survey

The GSS is conducted by the National Opinion Research Center (NORC) and was designed as part of a program of social indicator research to gather repeated measures on a broad range of data. The GSS used the NORC national probability sample, which includes all non-institutionalized English-speaking persons 18 years of age or older living in the United States. The samples were designed to give each household in the United States an equal probability of inclusion. Participants reported estimates of sexual partners via SAQ. The GSS asked questions about the number of sexual partners in the last year and since age 18 in 1989–1991, 1993, 1994, 1996, 1998, and 2000. Data on the number of partners the respondent had in the last year were also collected in 1988. There were a total of 16,159 participants who reported the number of sexual partners in the last year and 14,847 who reported the number of sexual partners since age 18 (Davis, Smith, & Marsden, 2003). GSS response codes for the number of sex partners in the last year were categorical and topcoded (1, 2, 3, 4, 5–10, 11–21, 21–100, 100+). As a result, when making comparisons between the GSS and other data sources for partners in the last year, we topcoded all studies at the highest common cutoff, 5+. GSS did not collect data on the number of partners in the lifetime, only partners since age 18, but one of the other data sets (NHSLS) collected data on partners since age 18 and we used this for comparison.

National College Health Risk Behavior Survey

The NCHRBS is a section of the Youth Risk Behavior Surveillance Survey conducted by the CDC in 1995 to monitor health-risk behavior among young college students. The NCHRBS used a two stage cluster sample design. The first-stage sampling frame contained 2,919 two and 4-year colleges and universities and the second stage was a random sample of individuals within the selected institutions. Of those eligible, 4,838 completed the questionnaire. Using a mailed SAQ, the NCHRBS asked about the number of male sexual partners and the number of female sexual partners over the lifetime. Each of these questions was top coded at 6. In order to get an estimate of the total number of sexual partners, we summed the responses to these two questions, but retained the topcode of six partners. The NCHRBS did not collect data on partners in the last year (CDC, 1997).

National Health and Social Life Survey

The NHSLS is a survey conducted in 1992 by the NORC. It was designed to be a comprehensive survey of the sexual behavior of adults 18–59 in the United States (Laumann, Gagnon, Michael, & Michaels, 1994). Participants were selected using a multistage area probability sample designed to give each household an equal probability of inclusion. A cross-sectional sample of 3,159 participants was collected as well as an over-sample of 273 black and Hispanic participants. The study used face-to-face interviews (FTFI) as well as SAQ to collect data on sexual experiences. The number of partners in the last year was asked multiple times in the survey using both modes of data collection. Laumann et al. constructed a categorical aggregate measure for partners in the last year that

was based on both the FTFI and SAQ data. For our comparative analysis of partners in the last year, we instead used the data exclusively from the FTFI in order to isolate the effects of data collection method and preserve the continuously coded responses. The number of lifetime partners for the NHSLS participants was not asked directly; instead, there was a set of questions on same sex and opposite sex partners before age 18, and a detailed set of life history calendar based questions on relationships after age 18. We constructed a lifetime measure following the strategy outlined by Laumann et al. in Appendix 5.2A, though again we excluded data from the SAQ to ensure a clean FTFI based measure. Data on partners since age 18 can be drawn from both the SAQ and FTFI sections of the survey independently allowing us to compare estimates from an SAQ and FTFI within the NHSLS as well as make a cross-survey comparison to the GSS.

National Survey of Family Growth

The NSFG is a multipurpose survey of a national sample of non-institutionalized women 15–44 years-old residing in the United States sponsored by the National Center for Health Statistics (U.S. Department of Health and Human Services, 1997). This analysis utilized data from cycle V of the survey, which was carried out in 1995 using FTFI. The sample for cycle V was a national probability sample of 10,847 women from households that had participated in the National Health Interview Survey (U.S. Department of Health and Human Services, 1997). The survey over-sampled both black and Hispanic women. The NSFG provides data on both the number of sexual partners in the last year and the number of sexual partners over the lifetime. For both questions, the instrument used by the NSFG allowed participants to give estimates in the form of a high and a low boundary if they could not recall an exact number. Only 0.4% (40) of participants could not recall an exact number of partners in the last year and 4.3% (468) of the participants could not recall an exact number for partners in their lifetime. Of the 468 participants reporting high and low estimates, 176 reported estimates that differed by only one partner. To reduce responses to a single variable, we took the mean from the high/low estimates and rounded to the nearest whole partner. There were also two cases reporting more sex partners in the last year than in their lifetime. To adjust for this discrepancy, lifetime partners were set equal to partners in the last year if partners in the last year were reported to be greater than partners in the lifetime.

National Survey of Men

The NSM was designed to examine sexual behavior and condom use among men. The study population consisted of 20–39 year old non-institutionalized men. The sample was based on a multi-stage, stratified, clustered, disproportionate-area probability sample of households within the contiguous United States and included an over-sample of Blacks. The data were collected in 1991 using FTFI, and the survey included questions for both partners in the last year and the lifetime (Tanfer, 1993). For both time frames, participants were asked to report the number of vaginal sex partners and anal sex partners. There is no way to ascertain how many partners were represented in both categories, so we defined the number of partners as the maximum of the two categories, which may be lower than the actual number of unique partners. A total of 586 (19%) of the men reported anal sex. Of these, 18 reported no vaginal sex partners and 35 reported more anal than vaginal sex partners. Finally, there were five

cases where the number of partners reported in the last year was greater than the number of partners reported for the lifetime, and for these we coded the lifetime partners equal to the last year.

National Survey of Women

The NSW was also conducted in 1991 and was designed to examine sexual, contraceptive, and fertility behaviors and the factors associated with these behaviors. The sample included 1,669 cases from two sub-samples. The first sub-sample ($n = 929$) consisted of follow-up cases from the 1983 National Survey of Unmarried Women, which surveyed 1,314 never-married women between 20 and 29 years of age. The second sub-sample ($n = 740$) was from a different probability sample of 20–27 year old women of unspecified marital status selected in 1991. Data were collected using FTFI {Tanfer, 1994 #63}. The NSW used a very similar instrument to the NSM, so the same adjustment strategy was used. About 17% of the women surveyed reported having had anal sex. If the number of partners reported in the last year was greater than the number of partners reported for lifetime, lifetime was set equal to the last year. There were only 14 such cases.

Behavioral Risk Factor Survey

The BRFS is a part of the state-based Behavioral Risk Factor Surveillance System initiated in 1984 by the Centers for Disease Control (CDC) to collect data on risk behaviors and preventative health practices (National Center for Chronic Disease Prevention and Health Promotion, 2003). The BRFS used telephone surveys and the questions regarding sexual behavior were part of a supplement started in 1996. States make the decision whether to include the supplement in each year. We used data for the 5 years from 1996 to 2000. The number of states that elected to include the supplement during that period varied from a high of 25 in 1997 to a low of 2 in 1996. The variation made it impossible to aggregate these data into a true national probability sample. The BRFS was also the only telephone survey in our comparison. We therefore ran all of the analyses with and without the BRFS in order to mitigate the impact it may have. The BRFS only included a question about partners in the last year so it was only included in a handful of our analyses. We did not want to exclude the BRFS entirely, however, because of the amount of data it provides ($n = 72,280$). The variable for the number of sex partners in the last year was topcoded at 76+. There were three participants reporting the topcoded value. As the number was similarly small for other surveys with continuous response coding, we topcoded all data for sex partners in the last year to 76+ in our analyses of the continuous response data.

Sample Composition Indicators—To account for differences in sample demographic composition, we used four primary demographic attributes: race, age, sex, and marital status. The demographic variables were recoded into categories dictated by lowest common denominator. For comparisons, race was collapsed to White, Black, and Other; marital status was collapsed to Married, Divorced-Widowed-Separated, and Never Married; and age was collapsed to 18–24, 25–34, and 35–44. Some of the surveys did not include participants of every age within each bin. The sample compositions for each study are shown in Appendix B.

These recoded demographic variables created 27 unique categories within sex. These categories created a demographic index which was used to control for basic differences in sample composition of the surveys. Differences in reported numbers of partners by the primary demographic attributes (sex, age, race, and marital status) were also used as a comparison metric for evaluating the differences observed by study.

Statistical Analysis—The analytic strategy we employed had two steps. The first step used exploratory data analysis to obtain a general picture of the aggregate differences in reporting across surveys. Data weighted by their original survey weights were used for this part of the analysis. The results represent what researchers would obtain if each study were analyzed separately, and what a simple comparison of published findings from these studies might show. In the second step, we use ANOVA to determine if the study or interview method had an independent effect on the reported number of partners after controlling for sample demographic composition. The data used for the ANOVA were weighted using the same post-stratification weights used in the exploratory data analysis with one small difference. To adjust for changes in the demographic composition of the sample population at the time the surveys were administered all of the data sets were adjusted to have the same demographic composition as the 2000 U.S. census.

The ANOVA model has indexed terms for demographics and study:

$$Y_{ijk} = \mu + d_i + s_j + \varepsilon_{ijk}$$

$i = (1, \dots, 26)$ demographic index category within sex,
 $j = (1, \dots, 6)$ adult study identifier,
 $k = (1, \dots, n_{ij})$ respondents in demographic category i of study j

Note that this is equivalent to a regression model with a set of dummy variables for demographic index and study.

The outcome variable was either the reported number of partners in the last year, the reported number of partners in the lifetime, or (for one specific analysis) the reported number of partners since age 18, and the analysis was performed separately by sex for each response coding of the outcome variable (continuous or top-coded). Each model was based on subsets of comparable data, and the comparability criteria are discussed below. Essentially, this analysis pools the datasets, cross-tabulates them by the demographic index (row) and study ID (column), and tests for study effects within demographic groups using the adjusted column sum of squares (SSQ), s_j in the equation above.

We focused on the SSQs and F -ratios from the ANOVA table because our analytic goal was to evaluate whether the variations in reporting between studies was “large.” The traditional ANOVA table output, decomposing the total SSQ into the contributions of study effects and demographic variation, was well suited to this question. ANOVA provided metrics for evaluating both the statistical significance and the substantive importance of the study effects.

Substantive importance was assessed in several ways: by assessing the magnitude of the adjusted study mean differences in percentage terms, by comparing the variance associated with the studies to the variance associated with the demographic attributes, and by graphical assessment of systematic discrepancies. When functions of the SSQ were compared, the sequence of terms entered into the model was demographics first and study ID last.

Determining the statistical significance of the observed differences across studies required a different approach than the traditional F -tests used in ANOVA and regression for variance components. Traditional tests compare the F -statistic observed in the data to an F -distribution with appropriate degrees of freedom, and the validity of the tests depends on a number of assumptions. Validity is threatened in our case by both the right skewed distribution of our response variable, which leads to inflated estimates of significance, and the complications induced by multiple survey designs. Statistical significance was therefore assessed using a permutation test—a resampling method like the bootstrap or jackknife. Resampling methods are widely used in statistics when the distribution of the sample test statistic is unknown. The specifics of the test are discussed below, but in essence, it replaces the standard reference F -distribution used to assess the probability of an observed F -statistic, with a “permutation distribution” of F -statistics generated from the data by randomly reassigning participants across studies (Good, 1994).

The permutation tested the hypothesis that the observed F -statistic for the study effect was larger than one would expect by chance, given the sample sizes and compositions of the studies. The F -statistic was first computed for the observed data. To construct the permutation distribution, observations within demographic strata were randomly assigned a study ID (permuted), the F -statistic was recomputed, and this process was repeated 1,000 times until a distribution of F -statistics was obtained. The p -value for the observed F -statistic can be estimated from this permutation distribution; it is simply the rank of the observed F divided by 1,000. This p -value tells us how likely the observed study differences are to have happened by chance, given the data in our studies. If all possible permutations were calculated, this would be an exact probability, but a full permutation would require an excessive number of computations in this case. The test remains exact to 1 divided by the number of permutations.

When the individual observations were permuted across study ID certain constraints were imposed: observations remained within their demographic category, and the number of observations in each study $\times$ demographic category cell for each permutation remained fixed. As a result, the permutation left many things unchanged: the study N 's, the sums of squares and df associated with the demographic index, and the df for the study effect and the residual variance. The permutation isolated and systematically varied only the sum of squares due to study ID and the residual sum of squares. The ratio of these mean squares in turn defined the F -statistic for study.

This permutation approach makes one assumption: the assumption of “exchangeability.” Individual observations are treated as though they could come from any study that included their demographic group. Put another way, this assumes our observations are independent and identically distributed (*iid*). Due to differences in sampling designs across studies, we

know these observations are not *iid*. Part of the difference is due to sample weights (due to both stratified sampling and post-stratification adjustment). We address this by creating self-weighting samples from each study. Self-weighted samples were generated by re-sampling observations with replacement from each study, with a probability defined by the study sample weights. The final self-weighted samples were constrained to have the same *N*, the same cell means for each demographic strata, and the same *F*-statistic for study ID as the original weighted samples. The other part of the difference is due to clustered sampling (used in some but not all of these studies). We did not control for that here, but the effect was to make our tests more conservative. By ignoring the clustering, our sample *N*s overstated the effective *N*s, and we were more likely to reject the null. In this case, accepting the null—that there were no significant differences between studies—was the finding of interest. So our test is conservative.

Several variations in study design made it necessary to restrict comparisons to subsets of similar studies. Three factors defined the comparable subsets: outcome measurement scale (continuous vs. topcoded responses), sample sex distribution (male only, female only), and data collection method (telephone, FTFI, SAQ). While only some of the studies were topcoded by design, continuous response data could be treated as topcoded, so we compared all surveys in the “topcoded” analysis, with the continuous responses topcoded to the lowest common denominator. For the subset of studies with continuous response coding for partners in the last year, the measure was actually topcoded at 76 for comparability with the BRFS. Some surveys were sex specific, so we conducted separate analyses for studies with males and studies with females. We also ran analyses excluding the BRFS, because it was the least likely to be representative and it had the largest sample (over 68,000 cases). The net result was 24 separate analyses across study subsets (3), outcomes (4), and sex (2).

Once the study effects in these comparable subsets were identified, we examined, where possible, whether the differences were associated with variations in interview method. For that analysis, the seven surveys were divided into FTFI, telephone, and SAQ groups. There were not enough replicates to provide a robust comparison of all three modes. The data did, however, permit a detailed within-study/between-method and within-method/between-study analysis of FTFI and SAQ for two surveys where the outcome response was partners since age 18.

Results

Comparisons of the unadjusted mean number of reported partners, and overall frequency distributions, showed clear but not large differences across the surveys. Figure 1 shows the results for partners in the last year and Fig. 2 shows the results for lifetime partners (with the exception of GSS, noted below). These figures represent what researchers would obtain if each study had been analyzed separately: that is, the data have been adjusted by their within-survey weights, but not adjusted for differences in sample composition across surveys.

The unadjusted frequency distributions shown in the figures were very similar, in some cases nearly identical. The subtle differences that did exist were enough to suggest statistically different overall means, as can be seen in the error bar plots. The nominal

confidence intervals in these figures assume a normally distributed response variable, however, so they do not accurately represent statistically significant differences.

One of the differences evident in the figures was the heaping on round numbers for lifetime partners. The only survey that did not show this was the NHSLS, and that was because the measure from this survey was based on summing a set of detailed life history calendar based questions, rather than one or two summary questions. Interestingly, however, the NHSLS distribution appeared to be a smoothed version of the heaped data from the other surveys, suggesting the two approaches did not produce fundamentally different results. The detailed NHSLS questions did elicit a slightly higher number of partners overall, however, for both men and women. This can be seen in the error bar plots, and this remained true after adjusting for sample composition in Table 2.

The means for the GSS are shown in the panels for lifetime partners for convenience, but they are based on partners since age 18, not the lifetime. The GSS means were expected to be lower than the other studies, and this is what was found. For partners in the last year, which were comparable across surveys, GSS was on the low side for Fig. 2b, but was in the middle or higher range once adjusted for sample composition in Table 2.

The means from the BRFS, which only has the measure for partners in the last year, were also consistently below the other surveys. In this case, however, there was no a priori reason to expect lower means. As noted above, the BRFS was the least likely to be representative of the U.S. population, and the only survey to use a phone interview.

For the most part, the differences in Table 2 were moderate. The largest differences were on the order of 0.3 partners for the year measure, and 0.9 (but more variable) for the lifetime measure. The single largest difference was two partners (about 13%), for the lifetime continuous measure for males, though there were only two studies with this measure. The SDs were 0.1 and 0.6 for the year and lifetime measures, respectively, and ranged from 3% to 15% of the overall averages.

After controlling for differences in sample composition, significant differences in reporting by study remained in six of the eight gender-by-outcome comparisons. These results are shown in Table 3. This table presents the results of the sequential progression through the data sources. In each step the data sets included were further restricted to surveys that had the most in common. We began by including all of the data, then the BRFS was excluded, and finally only studies that used FTFI were included.

Across all studies, there were no significant differences in the mean number of lifetime partners reported when the response was coded as a continuous variable, and this was true for both males and females. The observed F -statistic for these measures was in the center of the permutation distribution, the p -values were .58 and .31, respectively, so about what one would expect by chance.

For the remaining comparisons using all studies, the observed F -statistics were more than 40 times larger than the 5% threshold. When the BRFS was excluded, differences in male reports of partners in the last year were marginally significant for the remaining studies, with

p-values of .05 for the continuous response and .09 for the topcoded response. Significant study differences for the female reports persisted, though they were also smaller. Restricting the analysis to surveys that used FTFI did not change the significance of the remaining differences, suggesting that survey administration mode was not responsible for these significant study effects. By this point, however, only four of the eight outcome measure by sex differences were statistically significant below the .05 level.

The differences that remained after adjusting for sample composition were typically quite small, as suggested by Table 2. They were statistically significant because the amount of data in the studies used here gives the tests the power to detect even small differences in the means with a high certainty. It is worth considering other metrics, then, to evaluate whether the differences between studies are large or meaningful.

Table 4 compares the study effects to the main effects of age, race, and marital status, using the ratios of the MSE to adjust for differences in degrees of freedom. This ratio was well above 1 in 22 of the 24 comparisons. Across all the different outcome measures, comparison sets, and sex, the median effect of age was about 4 times larger than the study effect, the race effect was 3 times larger, and the marital status effect was 6 times larger. Excluding the BRFs strengthens the results: both instances of the study effect being larger than a demographic effect (marital status, for women) were reversed, and the relative sizes of the demographic effects were dramatically increased for men. In the FTFI-only comparison (which eliminates only GSS in the last year measure), the MSE ratios were all above 1 for both sexes. Note that, even though study differences remained statistically significant for women in this comparison, the median demographic effect was about 6 times larger. For men, the demographic effects were about 28 times larger than the residual study differences.

The study effects accounted for virtually none of the variance in partner reports. This can be seen at the end of the “All Studies” section at the top of Table 4. The fraction of the variance explained by study ranged from 0% to 1.3%, on the order of 10 times less than the variance associated with sample demographic composition. While the fraction was small for both, indicating substantial residual heterogeneity, demographics accounted for substantially more of the variance than study for all measures.

These results suggest a remarkable degree of consistency across studies and indicate that substantially more of the variance within and between studies was attributable to demographic characteristics.

Mode of Administration Effects

Adjusting for these demographic variations in sample composition helped to isolate the residual differences due to reporting, but the similarity in the results for the FTFI-only comparisons, and the comparisons with the other non-BRFs studies, suggested that the mode of survey administration may not account for much of this. This was somewhat unexpected, so we took a closer look to see if the finding was robust. The natural approach for testing the mode of administration effect would be to include a dummy variable in the ANOVA for mode and test it for significance, but there were not enough studies to provide sufficient replicates for a good test. Three studies used FTFI (NHSLs, NSM/NSW, NSFG),

but one of these only had female participants. Of the two studies that used SAQ (GSS and NCHRBBS), the former only had data on partners in the last year, and the latter only on lifetime partners, and both were topcoded. The BRFSS was the only survey to use telephone interviewing, and it only has data on partners in the last year. Overall, then, there was no robust statistical test possible across all three modes. But the patterns we did observe in the limited three-way comparisons available are worth describing, because the differences were not in the expected direction. Table 5 shows how often, within the 27 demographic subgroups for each sex, the means for the number of partners reported in the FTFI studies were higher than in surveys using the comparison mode of administration.

While one might have expected to see higher numbers of partners reported in the SAQ studies, especially for women, instead we found the opposite: SAQ studies in general produced lower numbers of reported partners. The one exception was among women, for partners in the last year, but even here the FTFI means were higher for about 20% of the demographic subgroups of women, so the SAQ advantage was not universal. The FTFI studies also consistently produced higher means than the one phone survey. Again, these differences were confounded with the effects of specific studies, and, as with all the study differences we observed, they were not large.

We can make one other comparison for FTFI and SAQ, for the outcome measuring the reported number of partners since age 18. The GSS collected data on partners since age 18 with an SAQ, and the NHSLs also collected data on this measure, with both an FTFI and SAQ. The SAQ question wording was nearly identical for the two studies (the wording is shown in Appendix A). This allowed for two effects to be estimated: a method effect (within study, across method) and a study effect (across studies, within method). The method effect was not significant in the permutation test ($p = .33$ and $p = .08$ for females and males, respectively). The study effect was not significant for females ($p = .47$), but was significant for males ($p < .01$). The joint distribution of the subgroup mean reports is plotted in Fig. 3. Within the NHSLs, the consistency in SAQ and FTFI reports was very strong, net of a few outliers for the males. Given that these were the same participants reporting in both modes, it might be expected that the reports would be close. Still, if the SAQ provided an opportunity for greater disclosure, there was no evidence that this encouraged participants to report something different. The bottom panels of the figure show the differences between the two studies when both used SAQ. The significant difference found for males was likely driven by the two unusually high group means for the NHSLs.

Response Coding Effects

Turning to response coding, the topcoded responses were consistently significantly different across surveys, even when the continuously coded responses were not. In the comparisons after removing the BRFSS, three of the four remaining significant differences were for topcoded responses. The topcoded analysis pooled all eligible studies: those in which the original responses were recorded with a topcode, and those for which we imposed a topcode post-facto (the different subsets of studies can be seen in Table 2). This may obscure a difference between original and post-facto top-coding. Topcoded response categories may influence the way people report partners, but we did not have enough replicates to test this.

In each comparison, we had only one study with an original topcode: GSS for the last year, and NCHRBS for the lifetime.

It may seem counterintuitive that topcoded responses were more likely to be significant than continuously coded responses, since top-coding eliminates the upper tail outliers that would be expected to influence statistical tests. Such outliers have little impact here, however, since the permutation test randomly reassigned them across studies for the permutation distribution. As a result, anomalous variations in the upper tail of the reporting distribution were less likely to induce a significant study effect than systematic variation in the lower tail, where there were many observations. By collapsing variation in the upper tail, topcoding magnified the relative impact of the differences at the lower tail (0 vs. 1). The result can be seen in Appendix C, which shows the fraction of participants that reported 0 and the topcode value for each measure. In general, differences in the 0 fraction were greater for women than for men, while differences in the topcode fraction were smaller. This corresponds to the patterns of statistical significance in Table 3: both of the top-coded comparisons were significant for women, and the one topcoded comparison significant for men had a study with an abnormally high 17% of men reporting 0 partners (NCHRBS, for partners in the lifetime).

Discussion

The number of sexual partners reported in the seven studies examined here initially appeared to be quite different, judging by the unadjusted means plotted in Figs. 1 and 2. On closer inspection, however, few of these differences turned out to be statistically significant or substantively large. We have eight comparisons for these studies, defined by measure (lifetime, last year), coding (continuous, topcoded), and sex of respondent. Overall, across these eight comparisons, study differences accounted for less than 1% of the variation in partner reports. Controlling for sample demographic composition, there were no significant study differences in the reported number of lifetime partners for either men or women when the measure was continuous. Topcoded measures of lifetime partners, however, were significantly different across studies for both sexes. For partners in the last year, there were significant study differences in both the continuous and topcoded measures for both sexes. A single study accounted for most of the difference in men's reports: the BRFs. This was also the only telephone survey and its sampling frame was the least likely to be representative. When we excluded the BRFs, study differences in partners reported in the last year fell to borderline significance for men, but remained significant for women.

Are these remaining significant differences important? The sample sizes ranged from 1,600 to 62,000, so our statistical tests were quite powerful, and we could detect differences of 0.2 of a partner/year as statistically significant. In this context, statistical significance is an excellent guide for inference, but may not be the best guide to substantive importance. The typical differences we observed between study means were on the order of 2–15%. Is this a large difference? By comparison, age, race, and marital status each had much larger effects: together, they accounted for 10 times more variation in partner reports and individually they accounted for a median of 4 times more per degree of freedom used. Compared to demographic effects, then, the study differences were small. Another metric for defining

substantive importance is the impact that these differences could have on predicted STI transmission. Here, the picture is less clear since the population level effects of average behavior differences can be highly nonlinear. Small differences in the mean can have a large impact on transmission dynamics, if these happen right at the reproductive threshold and raise R_0 above 1. An average difference of a single partner per year can do this (Morris & Dean, 1994), but recent simulation studies suggest that even differences as small as 0.2 of a partner can have this impact, especially in populations without the typical long right tailed partner distributions (Morris, Goodreau, & Moody, 2007). So, are the differences we find here important? The question should really be “important for what,” and the answer to this is, “it depends.” They are small in comparative magnitude, but the impact will depend on pathogen-specific transmission probabilities and how close the population level partnership network is to the reproductive threshold.

The factor most consistently associated with significant study differences was topcoding. This did not appear to be due to the influence a topcoded response category had on participants’ reports. We know this because one of our comparisons, lifetime partners in FTFI studies, used the same subset of studies in both the continuous and topcoded comparison. The survey methods, question formats, and even participants were identical in this case—the only parameter that varied was the post hoc imposition of a topcode. Yet, the continuous measure showed no significant differences between studies, while the topcoded measure produced significant study differences. Since the topcoded measure was less informative, the fact that it differed significantly across studies may be somewhat less important. It does, however, reflect real differences in the fraction of participants reporting 0 partners, a difference that is lost in the larger variation of the continuous measure. We did not have enough replicates to identify whether there were differences between responses originally collected with a topcoded measure and responses topcoded post hoc. Still, we would conclude from this finding that data on the number of sexual partners should be collected with continuous response coding. Topcoding does have the benefit of making surveys easier to keypunch or scan. But both topcoding and categorical coding create problems at the data analysis stage, especially for comparative analyses when the topcodes and categories may vary widely across studies. Both topcodes and categories can be imposed post hoc, so there is much to be lost by topcoding response categories and little to be gained.

The evidence also suggests that mode of administration had little systematic impact on the number of partners that participants report, though we did not have enough studies to conduct a robust statistical test. We found minimal differences between the FTFI and SAQ reports, suggesting that participants may be more willing to disclose this information than previously thought, and that the number of sexual partners may be less sensitive than behaviors like drug use and abortion. A closer look at the existing empirical literature supports this conjecture. The few studies that explicitly considered questions on the self-reporting of sexual partners were consistent with what we found here: there was either no difference in reporting associated with mode of administration (Boekeloo et al., 1994; Durant & Carey, 2000; Scandell, Klinkenberg, Hawkes, & Spriggs, 2003) or higher reporting in the FTFI (Solstad & Davidsen, 1993). The implication would seem to be that both modes produce consistent estimates, so the decision should be based on other grounds.

Overall, these findings indicate a remarkable level of consistency in the reporting of number of sexual partners across different subgroups of the U.S. adult population. With seven population-based studies of U.S. adults reviewed here, this finding has general implications for other similar studies. The limitations that should still be kept in mind are that the detailed results for mode of administration and response coding relied on fewer studies, and all of our studies were from the U.S. It would be unwise to infer beyond the U.S., since cultural norms may vary, and the resources to train interviewers may vary. But our findings provide clear evidence that within the U.S. differences across surveys are small. Consistency is not validity, but if systematic biases do exist, they must be uncannily consistent across 54 demographic subgroups, seven surveys, and at least two interview methods.

Given the differences in target populations, response rates, question wording, study purpose, mode of data collection, and survey organizations, this finding was somewhat surprising. We did not undertake the analysis expecting to find this outcome, and we used several different approaches to see if the results were robust: simple graphics (that stay close to the data), multiple metrics to capture formal and substantive definitions of a “large” difference, and robust statistical methods that did not depend on untenable assumptions. All of the approaches told the same story: the differences across studies were small in magnitude and generally not significant.

This is the first systematic comparison of self-reported numbers of sexual partners across a large group of population-based surveys. The consistency we found adds to a growing set of findings that challenge long-held views about the feasibility and accuracy of self-reported sexual behavior data. That said, it is worth reviewing some of this evidence, as both the public perception of self-reported sexual behavior, and the perception among other scientific fields, tend to cling to the stereotypes.

It is generally believed that people will refuse to answer questions on sexual behavior in surveys (Kolata, 2007; Lewontin, 1995; Smith, 1992). In fact, item non-response rates for sexual behavior questions are often extremely low. As shown in Table 1, 2% was typical, and 5% was the highest non-response for partner reports in all of the studies here. By contrast, item non-response rates for income questions are on the order of 25–30% in the Current Population Survey, which is routinely used as the basis for empirical research and policy recommendations.

There is also a general assumption that biomarkers provide the truth about sexual behavior while self-reports cannot be trusted. This belief ignores the errors generated by imperfect sensitivity and specificity of all biological tests. In large population based surveys (like Add Health, or NHANES), where prevalence of the target condition is low (for example, STDs), even tests with high levels of specificity will generate a large number of false positives. For example, with a sample of 14,000, and a true prevalence of 1%, a test with sensitivity of 94% and specificity of 99.5% will have a positive predictive value of only 65% (that is, 35% of the observed cases will be false positives). This is well understood by the bio-statisticians and epidemiologists who employ these tests, but is often ignored in popular discussion. Contrast 35% error to the 2–15% variation in study means for reported partners, and it is again puzzling why self-reported behavioral measures are regarded as more error prone.

Many continue to believe that there are large differences in men's and women's reports of the number of sexual partners. This too is a bit of a red herring. As shown in Morris (1993) using 3 years of the GSS data, almost all of the discrepancy in men's and women's means can be traced to reports in the top tail of the distribution. Among the 90% of participants reporting less than 20 lifetime partners, the ratio of male to female reports drops from 3.2:1 to 1.2:1. These differences may be statistically significant (large sample sizes ensure this), but they are much smaller than most people think.

Finally, it is often claimed that self-reports of sexual behavior are not concordant between partners. In fact, studies that enroll both partners find that reports for recent behaviors (frequency of coitus, types of sex, condom use) tend to be quite consistent, with rates of concordance 75% or higher (Bell, Montoya, & Atkinson, 2000; Harvey, Bird, Henderson, Beckman, & Huszti, 2004; James, Bignell, & Gillies, 1991; Lagarde, Enel, & Pison, 1995; Sison, Gillespie, & Foxman, 2004; Upchurch et al., 1991). With respect to reporting the existence of a sexual relationship, which was the focus of this article, the data suggest concordance may be higher still.

A recent paper shows concordance of over 90% in the reporting of sexual relationships among participants enrolled in a network study (Adams & Moody, 2007). In fact, sexual relationships were more likely to be reported by both partners than social, drug sharing, and needle sharing relations. Participants were also surprisingly accurate at reporting the existence of a sexual tie between two *other* people, with concordance rates of 95%. The participants in this study were a sample of drug injectors, prostitutes, and other persons at high risk, so not a group that one might expect to do well at a task like this.

This does not mean that there is no measurement error in self-reported sexual behavior, but it does suggest that the levels of error may be in the same range as many other survey measures. As is the case with all observational studies, the quality of the data depends on the quality of the study design, the training and performance of the field staff, and the care that has been taken in questionnaire development. It has been pointed out before that this is probably even more important for sexual behavior than for other self reported information (Cleland & Dare, 1994).

The results of this analysis are promising for research that relies on self-reported number of sexual partners. The small differences suggest that data on partner number can be collected effectively using a range of different survey methods and instruments. It also suggests that survey data on sexual behavior may be pooled across studies. Combining data from different sources provides powerful leverage for empirical research, providing larger samples for more detailed breakdowns, extending the range of sampled populations, and allowing detailed microdata to be linked to census or other aggregate population-based survey data (Handcock, Rendall, & Cheadle, 2005). Pooling may also provide a means by which we can efficiently investigate relatively rare events. Our findings suggest that the studies examined here can support a wide range of scientific inquiry in the future.

Appendix A

Table 6

Question wording and adjustments made to the data from seven surveys

Variable	Question for last year	Adjustments for last year	Question for lifetime	Adjustments for lifetime
BRFS	During the past 12 months, with how many people have you had sexual intercourse?	No adjustments. Responses topcoded at 76+. 3 participants reported 76+ partners	NA	NA
GSS	How many sex partners have you had in the last 12 months?	Responses were categorical for values greater than 4 so all responses greater than 4 were coded as 5. This study was treated as topcoded	Now thinking about the time since your 18th birthday (including the past 12 months), how many female partners have you had sex with? (same question asked for male partners)	Constructed by adding number of male partners since 18 and number of female partners since 18
NCHRB	NA	NA	During your life, with how many females have you had sexual intercourse? (same question asked for male partners)	Constructed by adding the number of male partners in lifetime and number of female partners in lifetime
NHSLS	Thinking back over the past 12 months, how many people, including men and women, have you had sexual activity with, even if only one time?	NA	Detailed questions based on a life history calendar. See Laumann et al. (1994) Appendix C for calculation NHSLS also includes explicit questions on partners since age 18: Now thinking about the time since your 18th birthday (again, including the recent past that you have already told us about), how many female partners have you ever had sex with? (same question asked for male partners)	For description, contact the corresponding author. Constructed by adding number of male partners since 18 and number of female partners since 18
NSFG	During the last 12 months, that is, since (month/year), how many men, if any, have you had sexual intercourse with? Please count every male sexual partner, even those you had sex with only once. (probe for range if R is unable to recall exact number)	Constructed from high-low estimates	Counting all your male sexual partners, even those you had intercourse with only once, how many men have you had sexual intercourse with in your life? (probe for range if R is unable to recall exact number)	Constructed from high-low estimates
NSM	Since January 1990, how many different women have you had vaginal intercourse with? Since January 1990, how many different partners have you had anal sex with?	Constructed from vaginal sex partners in 1990 and anal sex partners in 1990	With how many different women have you ever had vaginal intercourse? With how many partners have you ever had anal intercourse?	Constructed from vaginal sex partners in lifetime and anal sex partners in lifetime

Variable	Question for last year	Adjustments for last year	Question for lifetime	Adjustments for lifetime
NSW	With how many different men did you have vaginal intercourse since January 1990? With how many different men did you have vaginal intercourse since January 1990?	Constructed as the greater of vaginal sex partners in 1990 or anal sex partners in 1990	With how many different men have you ever had vaginal intercourse? With how many different men have you ever had anal intercourse?	Constructed from the greater of vaginal sex partners in lifetime or anal sex partners in lifetime

BRFS Behavioral Risk Factor Survey, *GSS* General Social Survey, *NCHRBS* National College Health Risk Behavior Survey, *NHSLs* National Health and Social Life Survey, *NSFG* National Survey of Family Growth, *NSM* National Survey of Men, *NSW* National Survey of Women

Appendix B

Table 7

Sample composition for each study

Adult studies	BRFS	GSS	NCHRS	NHLS	NSFG	NSM	NSW	Overall	2000 U.S. census
Age									
18–24	22.8	21.0	65.6	24.3	22.9	23.2	24.6	22.8	24.1
25–34	36.8	38.7	21.9	39.2	38.0	50.5	66.1	37.3	35.1
35–44	40.4	40.3	12.5	36.5	39.1	26.2	9.3	39.9	40.8
Race									
White	81.0	79.6	71.8	78.2	79.8	78.4	76.5	80.6	72.3
Black	11.4	13.4	10.2	11.8	13.7	11.6	16.4	12.3	12.8
Other	7.6	7.0	18.0	10.0	6.5	9.9	7.1	7.2	14.8
Sex									
Male	50.0	46.6	45.4	50.3	–	100.0	–	31.2	50.2
Female	50.0	53.4	54.6	49.7	100.0	–	100.0	68.8	49.7
Marital status									
Married	52.5	52.2	20.3	52.5	54.1	49.4	44.8	53.1	51.2
Divorced, widowed, separated	11.3	13.3	7.2	11.8	14.3	10.9	10.1	12.4	11.3
Never married	36.2	34.4	72.4	35.7	31.5	39.7	45.1	34.4	37.5
Total N	62,884	10,387	4,393	2,560	9,970	3,321	1,669	95,184	

Appendix C

Table 8

Proportion of observations reporting zero partners or the topcode value

Outcome: Sex:	Partners last year				Partners lifetime			
	Females		Males		Females		Males	
# of partners:	0	5+	0	5+	0	6+	0	6+
BRFS	15	0	13	3	NA	NA	NA	NA
GSS	21	1	13	3	5	19	5	43
NCHRB	NA	NA	NA	NA	13	31	17	38
NHSL	13	1	9	4	3	30	3	57
NSFG	18	1	NA	NA	11	25	NA	NA
NSW/NSM	29	1	8	7	7	29	6	53
Range	13–29	0–1	8–13	3–7	3–13	19–31	3–17	38–57

BRFS Behavioral Risk Factor Survey, GSS General Social Survey, NCHRB National College Health Risk Behavior Survey, NHSL National Health and Social Life Survey, NSFG National Survey of Family Growth, NSM National Survey of Men, NSW National Survey of Women

References

- Adams J, Moody J. To tell the truth: Measuring concordance in multiply-reported network data. *Social Networks*. 2007; 29:44–58.
- Bancroft, J., editor. *Researching sexual behavior: Methodological issues*. Indiana University Press; Bloomington, IN: 1997.
- Bell DC, Montoya ID, Atkinson JS. Partner concordance in reports of joint risk behaviors. *Journal of Acquired Immune Deficiency Syndromes*. 2000; 25:173–181. [PubMed: 11103048]
- Boekeloo BO, Schiavo L, Rabin DL, Conlon RT, Jordan CS, Mundt DJ. Self-reports of HIV risk factors by patients at a sexually transmitted disease clinic: Audio vs. written questionnaires. *American Journal of Public Health*. 1994; 84:754–760. [PubMed: 8179044]
- Bradburn, NM. Response effects.. In: Rossi, P.; Wright, J.; Anderson, A., editors. *Handbook of survey research*. Academic Press; New York: 1983. p. 289-328.
- Brody S. Patients misrepresenting their risk factors for AIDS. *International Journal of STD and AIDS*. 1995; 6:392–398. [PubMed: 8845395]
- Carey MP, Carey KB, Maisto SA, Gordon CM, Weinhardt LS. Assessing sexual risk behaviour with the Timeline Followback (TLFB) approach: Continued development and psychometric evaluation with psychiatric outpatients. *International Journal of STD and AIDS*. 2001; 12:365–375. [PubMed: 11368817]
- Catania JA, Binson D, van der Straten A, Stone V. Methodological research on sexual behavior in the AIDS era. *Annual Review of Sex Research*. 1995; 6:77–125.
- Catania JA, Gibson DR, Chitwood DD, Coates TJ. Methodological problems in AIDS behavioral-research: Influences on measurement error and participation bias in studies of sexual-behavior. *Psychological Bulletin*. 1990; 108:339–362. [PubMed: 2270232]
- CDC. Surveillance summaries. *MMWR*. 1997; 46(SS–6):1–54.
- Cleland JG, Dare OO. Reliability and validity of survey data on sexual behaviour. *Health Transition Review*. 1994; 4(Supp):93–110. [PubMed: 10150527]
- Davis, JA.; Smith, TW.; Marsden, PV. *General Social Surveys, 1972–2002: Cumulative codebook*. National Opinion Research Center, University of Chicago; 2003.
- Durant LE, Carey MP. Self-administered questionnaires versus face-to-face interviews in assessing sexual behavior in young women. *Archives of Sexual Behavior*. 2000; 29:309–322. [PubMed: 10948721]

- Durant LE, Carey MP. Reliability of retrospective self-reports of sexual and nonsexual health behaviors among women. *Journal of Sex and Marital Therapy*. 2002; 28:331–338. [PubMed: 12082671]
- Fenton KA, Johnson AM, McManus S, Erens B. Measuring sexual behavior: Methodological challenges in survey research. *Sexually Transmitted Infections*. 2001; 77:84–92. [PubMed: 11287683]
- Fu HS, Darroch JE, Henshaw SK, Kolb E. Measuring the extent of abortion underreporting in the 1995 National Survey of Family Growth. *Family Planning Perspectives*. 1998; 30:128–133. [PubMed: 9635261]
- Good, P. *Permutation tests: A practical guide to resampling methods for testing hypotheses*. Springer; New York: 1994.
- Gribble JN, Miller HG, Rogers SM, Turner CF. Interview mode and measurement of sexual behaviors: Methodological Issues. *Journal of Sex Research*. 1999; 36:16–24. [PubMed: 23226876]
- Handcock MS, Rendall MS, Cheadle JE. Improved regression estimation of a multivariate relationship with population data on the bivariate relationship. *Sociological Methodology*. 2005; 35:303–346.
- Harvey SM, Bird ST, Henderson JT, Beckman LJ, Huszti HC. He said, she said—Concordance between sexual partners. *Sexually Transmitted Diseases*. 2004; 31:185–191. [PubMed: 15076933]
- Huttenlocher J, Hedges LV, Bradburn NM. Reports of elapsed time: Bounding and rounding processes in estimation. *Journal of Experimental Psychology-Learning Memory and Cognition*. 1990; 16:196–213.
- James NJ, Bignell CJ, Gillies PA. The reliability of self-reported sexual behavior. *AIDS*. 1991; 5:333–336. [PubMed: 2059375]
- Jones R. Survey data collection using audio computer assisted self-interview. *Western Journal of Nursing Research*. 2003; 25:349–358. [PubMed: 12705116]
- Jones EF, Forrest JD. Underreporting of abortion in surveys of United States women: 1976 to 1988. *Demography*. 1992; 29:113–126. [PubMed: 1547898]
- Kolata, G. The myth, the math, the sex. *The New York Times*; Aug 12. 2007 Retrieved from <http://www.nytimes.com/2007/08/12/weekinreview/12kolata.html>
- Lagarde E, Enel C, Pison G. Reliability of reports of sexual behavior: A study of married couples in rural West Africa. *American Journal of Epidemiology*. 1995; 141:1194–1200. [PubMed: 7771458]
- Laumann, EO.; Gagnon, JH.; Michael, RT.; Michaels, S. *The social organization of sexuality: Sexual practices in the United States*. University of Chicago Press; Chicago: 1994.
- Lewontin, RC. Sex, lies, and social science. Vol. 42. *The New York Review of Books*; 1995.
- McLaws M, Oldenburg B, Ross MW, Cooper DA. Sexual behavior in AIDS-related research: Reliability and validity of recall and diary measures. *Journal of Sex Research*. 1990; 27:265–281.
- Metzger DS, Koblin B, Turner C, Navaline H, Valenti F, Holte S, et al. Randomized controlled trial of audio computer-assisted self-interviewing: Utility and acceptability in longitudinal studies. *American Journal of Epidemiology*. 2000; 152:99–106. [PubMed: 10909945]
- Morris M. Telling tails explain the discrepancy in sexual partner reports. *Nature*. 1993; 365:437–440. [PubMed: 8413586]
- Morris M, Dean L. The effects of sexual behavior change on long-term HIV seroprevalence among homosexual men. *American Journal of Epidemiology*. 1994; 140:217–232. [PubMed: 8030625]
- Morris, M.; Goodreau, S.; Moody, J. Sexual networks, concurrency, and STD/HIV.. In: Holmes, KK.; Sparling, PF.; Mardh, PA.; Lemon, SM.; Stamm, WE.; Piot, P.; Wasserheit, JN., editors. *Sexually transmitted diseases*. McGraw-Hill; New York: 2007. p. 109-125.
- National Center for Chronic Disease Prevention and Health Promotion. BRFSS User's Guide. Aug 11. 2003 from http://www.cdc.gov/brfss/technical_infodata/usersguide.htm
- Saltzman SP, Stoddard AM, McCusker J, Moon MW, Mayer KH. Reliability of self-reported sexual-behavior risk-factors for HIV-infection in homosexual men. *Public Health Reports*. 1987; 102:692–697. [PubMed: 3120234]
- Scandell DJ, Klinkenberg WD, Hawkes MC, Spriggs LS. The assessment of high-risk sexual behavior and self-presentation concerns. *Research on Social Work Practice*. 2003; 13:119–141.

- Schwarz N, Strack F, Mai HP. Assimilation and contrast effects in part-whole question sequences—A conversational logic analysis. *Public Opinion Quarterly*. 1991; 55:3–23.
- Sison JD, Gillespie B, Foxman B. Consistency of self-reported sexual behavior and condom use among current sex partners. *Sexually Transmitted Diseases*. 2004; 31:278–282. [PubMed: 15107629]
- Smith TW. Discrepancies between men and women in reporting number of sexual partners: A summary from 4 countries. *Social Biology*. 1992; 39:203–211. [PubMed: 1340040]
- Sneed CD, Chin D, Rotheram-Borus MJ, Milburn NG, Murphy DA, Corby N, et al. Test-retest reliability for self-reports of sexual behavior among Thai and Korean participants. *AIDS Education and Prevention*. 2001; 13:302–310. [PubMed: 11565590]
- Sohler N, Colson PW, Meyer-Bahlburg HFL, Susser E. Reliability of self-reports about sexual risk behavior for HIV among homeless men with severe mental illness. *Psychiatric Services*. 2000; 51:814–816. [PubMed: 10828118]
- Solstad K, Davidsen M. Sexual-behavior and attitudes of Danish middle-aged men: Methodological considerations. *Maturitas*. 1993; 17:139–149. [PubMed: 8231905]
- Tanfer K. National survey of men: Design and execution. *Family Planning Perspectives*. 1993; 25:83–86. [PubMed: 8491290]
- Tanfer K. 1991 National Survey of Women. 1994 Battelle Memorial Institute, Centers for Public Health Research and Evaluation (Producer) Sociometrics Corporation, AIDS/STD Data Archive (Producer & Distributor) Seattle, WA Los Altos, CA (Data Set 17–19, McKean, E. A., Muller, K. L., & Lang, E. L.) [machine-readable data file and documentation].
- Taylor JF, Rosen RC, Lieblum SL. Self-report assessment of female sexual function: Psychometric evaluation of the Brief Index of Sexual Functioning for Women. *Archives of Sexual Behavior*. 1994; 23:627–643. [PubMed: 7872859]
- Tourangeau R, Rasinski KA. Cognitive-processes underlying context effects in attitude measurement. *Psychological Bulletin*. 1988; 103:299–314.
- Tourangeau R, Rasinski KA, Bradburn N. Measuring happiness in surveys: A test of the subtraction hypothesis. *Public Opinion Quarterly*. 1991; 55:255–266.
- Tourangeau R, Smith TW. Asking sensitive questions: The impact of data collection mode, question format, and question context. *Public Opinion Quarterly*. 1996; 60:275–304.
- Turner CF, Ku L, Rogers SM, Lindberg LD, Pleck JH, Sonenstein FL. Adolescent sexual behavior, drug use, and violence: Increased reporting with computer survey technology. *Science*. 1998; 280:867–873. [PubMed: 9572724]
- Upchurch DM, Weisman CS, Shepherd M, Brookmeyer R, Fox R, Celentano DD, et al. Interpartner reliability of reporting of recent sexual behaviors. *American Journal of Epidemiology*. 1991; 134:1159–1166. [PubMed: 1746526]
- U.S. Department of Health, Human Services. National survey of family growth, Cycle V, 1995. U.S. Department of Health and Human Services, National Center for Health Statistics; Hyattsville, MD: 1997.
- Van Duynhoven YTHP, Nagelkerke NJD, Van de Laar MJW. Reliability of self-reported sexual histories: Test-retest and interpartner comparison in a sexually transmitted diseases clinic. *Sexually Transmitted Diseases*. 1999; 26:33–42. [PubMed: 9918321]
- Weinhardt LS, Carey MP, Maisto SA, Carey KB, Cohen MM, Wickramasinghe SM. Reliability of the timeline follow-back sexual behavior interview. *Annals of Behavioral Medicine*. 1998a; 20:25–30. [PubMed: 9755348]
- Weinhardt LS, Forsyth AD, Carey MP, Jaworski BC, Durant LE. Reliability and validity of self-report measures of HIV-related sexual behavior: Progress since 1990 and recommendations for research and practice. *Archives of Sexual Behavior*. 1998b; 27:155–180. [PubMed: 9562899]

Fig. 1. Reporting number of partners last year by study, sex and response coding. **a** (continuous), **b** (topcoded). *Note:* Data weighted by original study weights, but not adjusted for differences in sample composition. Frequency distributions are shown on the left, means and nominal confidence intervals on the right

Fig. 2.

Reported number of partners in the lifetime by study, sex and response coding. **a** (continuous), **b** (topcoded). *Note:* GSS reports partners since age 18, not lifetime. Data weighted by original study weights, but not adjusted for differences in sample composition. Frequency distributions are shown on the left, means and nominal confidence intervals on the right

Fig. 3. Effect of method and study on reported number of partners since age 18. *Note:* Each point represents the mean for a demographic group defined by race, age, marital status and sex

Table 1

Overview of the seven population based surveys

Survey	Years	Age	Sex	Interview method	Survey Resp. Rate (%)	Eligible sample	Partners in last year		Partners in lifetime	
							N (coding)	Item Resp. Rate	N (coding)	Item Resp. Rate
BRFS	1996–2000	18+	M/F	Phone	49–63	62,884	59,927	95.3	NA	
GSS ^a	1988–2000	18+	M/F	SAQ	70–82	10,387	9,151 (T)	99.0	8,735	97.2
NCHRS	1995	18+	M/F	SAQ	60	4,393	NA		4,328 (T)	98.5
NHSLS	1992	18–59	M/F	FTFI and SAQ	79	2,560	2,554	99.8	2,559	100.0
NSFG	1995	15–44	F	FTFI	79	9,970	9,900	99.3	9,776	98.1
NSM	1991	19–41	M	FTFI	70	3,321	3,320	100.0	3,317	99.9
NSW	1991	19–38	F	FTFI	71	1,669	1,669	100.0	1,669	100.0

Notes: All continuous year data were topcoded at 76+ to match the BRFS. For other measures, the (T) for (coding) indicates the original response was recorded with a topcode. Eligible sample size was based on the number of participants who were asked the sexual behavior questions, were in the target age range, and were not missing their demographic information. Item response rates based on refusals and “no answers” in direct and filter questions where appropriate

^aThe GSS collected data on partners since age 18, in 1988, 1991, 1993, 1994, 1996, 1998, and 2000. It is analyzed separately below. See Appendix A for details

Table 2

Study means for reported number of partners weighted to 2000 U.S. census demographics by sex and outcome measure

Outcome measure		Females				Males			
		Last year		Lifetime		Last year		Lifetime	
		C	T	C	T	C	T	C	T
Survey	BRFS	0.96	0.95			1.33	1.20		
	GSS		1.10				1.40		
	NCHRBS				3.70				4.18
	NHSLS	1.13	1.10	5.96	3.63	1.63	1.40	17.24	4.46
	NSFG	1.04	1.02	5.61	3.20				
	NSW/NSM	0.86	0.84	6.49	3.51	1.82	1.47	15.15	4.31
Mean of study means		1.0	1.0	6.0	3.5	1.6	1.4	16.2	4.3
Largest difference		0.27	0.27	0.88	0.50	0.49	0.27	2.09	0.28
SD between studies		0.12	0.11	0.44	0.22	0.25	0.12	1.48	0.14
SD/mean		12%	11%	7%	6%	15%	8%	9%	3%

Note: Bottom row shows the standard deviation between studies divided by the mean of the study means

C Continuous response coding, *T* Topcoded responses

Table 3

Study effect *p*-values from the permutation tests

		Females				Males			
		Last year		Lifetime		Last year		Lifetime	
		C	T	C	T	C	T	C	T
All studies	<i>p</i> -value	<.001	<.001	BRFS, NHSLs, NSFG, NSW	BRFS, GSS, NHSLs, NSFG, NSW	<.001	<.001	.31	<.001
	Studies compared	BRFS, NHSLs, NSFG, NSW	BRFS, GSS, NHSLs, NSFG, NSW	NHSLs, NSFG, NSW	NCHRBS, NHSLs, NSFG, NSW	BRFS, NHSLs, NSM, NSW	BRFS, GSS, NHSLs, NSM	NHSLs, NSM, NSW	NCHRBS, NHSLs, NSM
Excluding BRFS	<i>p</i> -value	<.001	<.001			.05	.09		
	Studies compared	NHSLs, NSFG, NSW	GSS, NHSLs, NSFG, NSW	As above	As above	NHSLs, NSM, NSW	GSS, NHSLs, NSM	As above	As above
FTFI studies only	<i>p</i> -value		<.001				.14		.02
	Studies compared	As above	NHSLs, NSFG, NSW	As above	NHSLs, NSFG, NSW	As above	NHSLs, NSM	As above	NHSLs, NSM

C Continuous response coding, T Topcoded responses

Table 4

Relative effects of demographic attributes versus study effect using MSE ratios

Data sets	Comparison attribute	Females				Males			
		Year		Life		Year		Life	
		C	T	C	T	C	T	C	T
All studies	Age	4.8	3.0	68.4	6.6	1.6	3.3	12.3	3.8
	Race	1.2	1.2	16.3	3.2	1.4	3.2	9.9	3.0
	Marital status	0.7	0.9	313.5	10.3	3.8	7.5	40.4	1.2
<i>Percent of variance due to:</i>	<i>Sample composition</i>	<i>1.9%</i>	<i>33%</i>	<i>3.9%</i>	<i>8.9%</i>	<i>3.7%</i>	<i>7.8%</i>	<i>5.0%</i>	<i>12.9%</i>
	<i>Study</i>	<i>0.2%</i>	<i>0.7%</i>	<i>0%</i>	<i>0.6%</i>	<i>0.5%</i>	<i>0.8%</i>	<i>0%</i>	<i>1.3%</i>
Excluding BRFS	Age	6.4	2.7	As above	As above	4.2	56.9	As above	As above
	Race	3.7	1.5			9.6	33.0		
	Marital status	5.8	3.9			24.4	122.0		
FTFI studies only	Age	As above	2.7	As above	5.2	As above	46.5	As above	13.7
	Race		1.3		3.2		32.7		35.8
	Marital status		2.7		9.9		80.8		46.0

Note: Cells show ratio of demographic attribute MSE to study MSE for each attribute

C Continuous response coding, *T* Topcoded responses

Variance attributable to sample composition and study are shown in italics

Table 5

FTFI reports compared to other modes

Comparison mode	Comparison survey	Measure	How often FTFI study means are higher	
			Males (%)	Females (%)
SAQ	NCHRBBS only	Life, topcoded	63	19
	GSS only	Year, topcoded	78	56
Phone	BRFS only	Year	89	70

Note: FTFI surveys include NHSLs, NSFG and NSM/W, as appropriate for each sex. Percentages based on comparing cell means across modes within each demographic index category